

# 深層学習に基づく楽器音分類のための画像分類ネットワークを用いた ファインチューニング

城間 佑樹<sup>†</sup> 木下 裕磨<sup>†</sup> 塩田さやか<sup>†</sup> 貴家 仁志<sup>†</sup>

<sup>†</sup> 東京都立大学大学院システムデザイン研究科情報科学域  
〒191-0065 東京都日野市旭が丘 6-6

**あらまし** 本研究では、楽器音分類タスクに画像分類ネットワークを用いる際のファインチューニングのためのチャンネル変換法の比較評価を行う。近年、環境音識別や感情認識など様々なタスクにおいて深層学習を用いた手法が提案されている。また、深層学習に用いる学習データが少ない場合に、画像分類タスクのネットワークでファインチューニングを行うことで音に関するタスクの性能が改善することが報告されている。音を入力とする場合は画像ネットワークの入力に対応させるためにスペクトログラムを用いることが多いが、スペクトログラムが1チャンネルデータとなっているのに対し、画像の入力を前提にしたネットワークはRGBの3チャンネルデータが入力されることを想定しているため、チャンネル数を合わせる必要がある。チャンネルを変換する手法としてこれまでに、各チャンネルに同じデータを複製をする手法、動的特徴量を用いる手法、スペクトログラムをカラー画像化する手法などが提案されているが、手法の違いが精度にどの程度影響するか明らかにされていない。そこで本研究では、様々なチャンネル変換法がファインチューニングの結果にどのような影響を与えるのかについて比較を行う。本実験では、ImageNetと呼ばれる大規模な画像データを用いて学習されたネットワークに対してファインチューニングを行い楽器音分類を行った。チャンネル変換法として6種類の手法を比較したところ、実験結果よりカラー画像化がImageNetに適していたことを報告する。

**キーワード** 楽器音分類, 画像分類ネットワーク, ファインチューニング, チャンネル変換

## Investigation on fine-tuning with image classification networks for deep neural network-based musical instrument classification

Yuki SHIROMA<sup>†</sup>, Yuma KINOSHITA<sup>†</sup>, Sayaka SHIOTA<sup>†</sup>, and Hitoshi KIYA<sup>†</sup>

<sup>†</sup> Department of Computer Science, Graduate School of Systems Design, Tokyo Metropolitan University  
6-6 Asahigaoka, Hino-shi, Tokyo, 191-0065 Japan

**Abstract** In this paper, we investigate abilities of channel conversion methods for fine-tuning with image classification networks under deep neural network-based musical instrument classification. Recently, many deep neural network-based methods have been proposed for scene classification, emotion recognition tasks, and so on. It has also been reported that fine-tuning techniques with well-trained networks using large-scale image dataset improve the performance of sound classification tasks when the limited amount of training data is available. In this case, while a spectrogram extracted from sound data is usually regarded as an image and inputted to the fine-tuned networks with the image classification tasks, the spectrogram image is not suitable to the fine-tuned network because the input of the image classification networks assumes the three channel data like RGB. In this case, the spectrogram is required to be converted to the three channel data, and many methods such as spectrogram duplication method, a method using delta as coefficients and colorization of a spectrogram have been proposed. However, there is no discussion how these methods affect the accuracies. Therefore, we compare various channel conversion methods via fine-tuning of the image classification networks. In the experiments, we performed musical instrument classification with fine-tuning of the well-trained networks by ImageNet. From the results, compared among six channel conversion methods, the colorization of a spectrogram was the most suitable for the fine-tuning with the image classification networks.

**Key words** Acoustic musical instrument classification, image classification network, fine-tuning, channel conversion

## 1. まえがき

近年、深層学習の発展により、音声だけでなく環境音や楽器音など様々な音の分類に関するタスクが活発に研究されるようになってきている [1]. 対象とする音は様々あり、空港やショッピングモールといった録音環境を予測する環境音分類タスク [2,3], 食器の音や掃除機の音など家庭内の音を予測する家庭内音分類タスク [4,5], 演奏された楽器の音からそれがどの楽器のものか予測する楽器音分類タスク [6,7] などが存在している. これらのタスクのために近年, Convolutional Neural Network (CNN) に基づく手法が多く提案されている [8–10]. これらの手法では入力音からスペクトログラムを抽出し, CNN に入力して分類を行っている.

一般的に深層学習に基づく手法において精度をあげるためには大量の学習データ量が必要となることが知られている. そのため, データが確保できない場合に, 高い精度を得るための手法の一つとしてファインチューニングが有効であることが知られている [11]. ファインチューニングは目的のタスクで使用するデータセットとは異なるタスクのための巨大なデータセットを用いてネットワークを事前学習し, その重みを初期重みとして目的のタスクで使用するデータセットを用いてネットワーク学習するというものである. ファインチューニングは画像分類の分野で大きな成功を収めており, 特に 2009 年に登場した ImageNet [12] と呼ばれる巨大なデータセットで事前学習されたネットワークを用いることで多くの画像分類タスクで高い精度を達成したことが報告されている [13,14]. さらに, ImageNet を使ったファインチューニングは画像分類タスクだけでなく音を使ったタスクである環境音分類や感情認識などのタスクにおいても有用であることが報告されている [15–19].

画像データで事前学習されたモデルを音響タスクに使用する場合, モデルの入力チャンネル数とスペクトログラムのチャンネル数が一致しないという問題が発生する. これは, スペクトログラムがスペクトルの振幅値に従って色付けしたものであり実際には 1 チャンネルデータである一方, 画像データで事前学習されたモデルは入力画像の R,G,B の 3 チャンネルで構成されているためである. これまでの研究ではこの問題に対処するために, 3 チャンネルすべてに同じデータを複製して入力する手法 [15] や, スペクトログラムの時間方向の 1 次及び 2 次動的特徴量をとる手法 [16], スペクトログラムをカラー画像化する手法 [17] などが提案されている. しかし, これまでにこれらデータの入力方法の違いがシステムの性能にどの程度影響するか明らかになっていない.

そこで本研究では, チャンネル変換手法の違いが音の分類タスクの性能にどのような影響を与えるのかを評価するために, 直接的なチャンネル変換法と間接的なモデルベースのチャンネル変換法について楽器音分類タスクを用いて性能を調査した. 本実験では, ResNet50 [20] および VGG16 [21] という画像分類で大きな成果を上げた 2 つのモデルを用い, それぞれファインチューニングの有無で評価を行った. データベースには NSynth [22] と FSDKaggle2018 [23] のそれぞれから取り出した単音の楽器

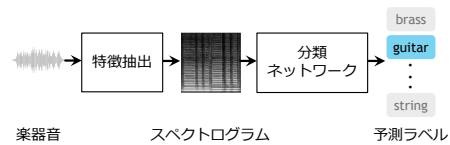


図1 楽器音分類のフロー図

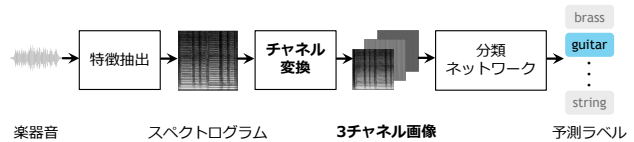


図2 画像ネットワークを用いた楽器音分類のフロー図

音及び演奏音を用いた. チャンネル変換法には 6 種類の手法を用いて評価したところ, 実験結果より, カラー画像化が画像分類ネットワークのファインチューニングに適していることを報告する.

## 2. 楽器音分類タスクとファインチューニング

楽器音分類とは, 入力された楽器音がどの楽器の音であるか予測するというタスクである. 近年提案されている CNN に基づく手法では図 1 に示すフローのように, 入力された楽器音からスペクトログラムを抽出し, そのスペクトログラムをネットワークに入力することで予測結果を得ている. 評価においては分類ネットワークの最終層を softmax 層とすることで最も確率値の高い楽器を予測結果としている.

深層学習を用いる手法において, 一般的に大規模な学習データが高い性能を得るために必要であることが知られているが, 使用可能なデータが小規模なものに限られる場合には, データ量不足を補うためにファインチューニングを行うことがある [24]. ファインチューニングを行う際はネットワーク構造を維持するために学習元のタスクと同じようにデータを入力する必要がある. また, 元のタスクと目的のタスクとのドメインの違いや正規化などについても注意が必要となる.

## 3. ファインチューニングのためのチャンネル変換

画像分類タスクのために学習されたネットワークに対してファインチューニングを行い, 楽器音分類を行うことを考える. 画像ネットワークは入力データの値のレンジが 0 から 255 であることを想定している. そのため, 楽器音のスペクトログラムを入力する際は, スペクトログラムの値を 0 から 255 に収まるよう正規化を行う, グレースケール画像化を行っている. しかし, スペクトログラムをグレースケール化したとしてもチャンネル数は依然 1 つである. そのため, カラー画像で学習されたネットワークの入力として想定されている画像の R,G,B チャンネルと合うように入力をチャンネル変換して 3 チャンネルに合わせる必要がある. このチャンネル変換を含む楽器音分類のフローを図 2 に示す. 入力された楽器音からスペクトログラムを抽出す

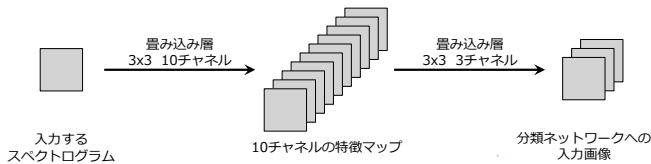


図3 手法 (e) での畳み込み層 (2 層)

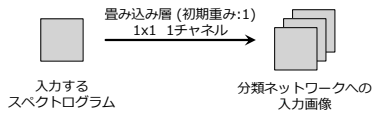


図4 手法 (f) での畳み込み層 (1 層)

るのは図1と同様だが、その後にはチャンネル変換を介することで3チャンネルデータに変換してネットワークに入力するという流れになっている。

### 3.1 チャンネル変換手法

画像分類ネットワークを用いたファインチューニングを行う際に使われている主なチャンネル変換法について述べる。

#### (a) 複製

複製は3チャンネルすべてに同じグレースケール化したスペクトログラムを用いたものである [15]。R,G,B すべて全く同じ値となるため、元の ImageNet の入力とは想定が異なるが、R,G,B の相関があまり強くない場合には親和性が高い方法だと考えられる。ただし、同じ情報を3回入力するという冗長性がある。

#### (b) 動的特徴量

音の特徴量抽出では時間差分を動的特徴量として埋め込む手法が広く用いられている。動的特徴量を考慮したチャンネル変換法では、1チャンネル目にグレースケール化したスペクトログラム、2チャンネル目に1次動的特徴量、3チャンネル目に2次動的特徴量を入力している [16]。時系列データである音にとって動的特徴量は重要である一方、画像では時間差分を考慮するようなチャンネル間の相関が適切に学習可能かに性能が依存すると考えられる。

#### (c) カラー画像化

スペクトログラムの値に対応するカラーチャートを設定することでカラー表現したスペクトログラムをカラー画像とみなし、R,G,B の3チャンネルに分けて入力する手法である [17]。カラー画像化のカラーパターンに自由度があるがチャンネル変換後の画像はファインチューニングに使うネットワークと親和性が高いと期待できる。

#### (d) 帯域分割

スペクトログラムをグレースケール画像とみなし、それを周波数の低域・中域・高域の3つの帯域に分割することで3チャンネルデータを作成するものである [18]。他の手法はスペクトログラムを正方形の画像と捉えて入力を行っているが、帯域分割によるチャンネル変換では周波数方向に3分割するために元のスペクトログラムの形状が縦長で、そこから帯域ごとに分割する

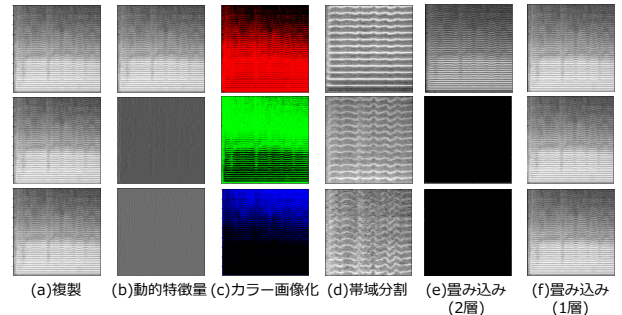


図5 各チャンネル変換法で生成した3チャンネルの画像 (NSynth, 'brass' のデータ)

ことを行っている。

#### (e) 畳み込み (2 層)

入力のスペクトログラムと分類ネットワークの間を繋ぐように畳み込み層を挟むことで1チャンネルから3チャンネル入力に変換する手法である [19]。文献 [19] では図3に示すように1チャンネルのスペクトログラムを最初の畳み込み層で10チャンネルを増やし、次の畳み込み層で3チャンネルにしている。分類ネットワークとは2層目の出力から繋がるよう学習が行われる。1チャンネルのグレースケール化したスペクトログラムから3チャンネルへの変換そのものを学習できるため、よりネットワークに適切な特徴量抽出が可能になることが期待されている。ただし、カーネルサイズを3x3、重みの初期値をGlorotの正規分布で初期化しているため学習データが少ない場合に学習が難しくなる場合がある。

#### (f) 畳み込み (1 層)

手法 (e) と同様に畳み込み層により1チャンネルから3チャンネルに変換する手法である。図4に手法 (f) のフロー図を示す。手法 (e) との違いは、畳み込み層のカーネルサイズを1x1、重みの初期値を1にすることで、初期状態を手法 (a) と同じグレースケール化したスペクトログラムを複製する状態にしている点である。手法 (e) では初期重みがランダムとなっているが、手法 (f) では初期状態を手法 (a) と同じ状態にしているため、学習データが少ない場合にも安定してチャンネル変換法を学習できることが期待できる。

手法 (a) から (d) が1枚の画像から3枚の画像への直接的なチャンネル変換法となっており、手法 (e),(f) がモデルベースの間接的なチャンネル変換法となっている。

### 3.2 生成画像の比較

図5に各チャンネル変換法によって生成された3チャンネル分の画像を示す。上から1チャンネル目、2チャンネル目、3チャンネル目の画像となっている。手法 (a) は同じスペクトログラムを複製しているだけなので、3チャンネルが全て同じ入力画像となっている。手法 (b) は分かりづらくなっているが1チャンネル目の画像のエッジが2チャンネル目の画像に、2チャンネル目の画像のエッジが3チャンネル目の画像に表れている。手法 (c) はスペクトログラムの値の低いものから高いものにかけて色相が青から赤に変化するように着色するため、Python 標準ライブラリであ

表1 NSynth を用いた分類正解率 (%)

	モデル	ResNet50		VGG16		
		Fine-Tuning の有無	なし	あり	なし	あり
直接	(a) 複製		78.1	<b>86.5</b>	78.9	<b>83.1</b>
	(b) 動的特徴量		<b>65.6</b>	62.5	81.7	<b>83.3</b>
	(c) カラー画像化		83.0	<b>87.8</b>	81.4	<b>83.1</b>
	(d) 帯域分割		79.7	<b>86.2</b>	78.4	<b>85.1</b>
間接	(e) 畳み込み (2 層)		77.7	<b>85.6</b>	76.6	<b>79.0</b>
	(f) 畳み込み (1 層)		79.6	<b>85.9</b>	78.1	<b>78.4</b>

る Matplotlib を用いてカラーマップを設定することでスペクトログラムのカラー画像を作成した。わかりやすさのため、それぞれのチャンネルが意味する色に合わせて表示しているが実際にはグレースケールのようにそれぞれ 0 から 255 の値を持つ画像で入力される。手法 (d) は周波数方向のサイズがモデルの入力形状の 3 倍の大きさを持つスペクトログラムを生成し、それを低域・中域・広域の 3 つに等分割された画像になっている。周波数方向に 3 分割するため、他の画像と異なり周波数の解像度が高くなっており、調波構造の間隔が他よりも広く明確に表示されていることがわかる。手法 (e) 及び (f) は、モデルを学習した後の畳み込み層が出力する特徴マップを表示している。手法 (e) は 1 チャンネル目がスペクトログラムと同じような見た目をしており、2 チャンネル目及び 3 チャンネル目は画像内の殆どの値が 0 の黒い画像となっている。手法 (f) は全チャンネルがスペクトログラムと同じような見た目になっている。ただし複製とは異なり 3 チャンネルのデータは全く同じにはなっていない。

## 4. 楽器音分類実験

### 4.1 データセット

本実験では、楽器音のデータベースとして NSynth [22] と FSDKaggle2018 [23] の一部を使用した。データ量を均一にするために各データセットからそれぞれ 4 種類の楽器を選択して実験を行っている。NSynth では音源の情報として Acoustic 及び Electric のラベルが付いているものの中から brass, guitar, keyboard, string のラベルがついているもののみを選択した。各楽器それぞれ学習用データ数を 500 サンプル、評価用データ数を 800 サンプル、テスト用データ数を 265 サンプルとした。FSDKaggle2018 では、人手でつけられたラベルのうち acoustic guitar, cello, electric piano, trumpet のラベルが付いている 4 楽器を選び、それらを 2 秒の長さごとに分割しそれぞれ 1 サンプルとした。各楽器それぞれの学習用データ数を 130 サンプル、評価用データ数を 29 サンプル、テスト用データ数を 80 サンプルとした。NSynth は 16 kHz, FSDKaggle2018 は 22.5 kHz のサンプリング周波数である。

### 4.2 実験条件

チャンネル変換法には入力音波形から抽出した対数振幅スペクトログラムを入力している。フレーム長は 512 点、フレームシフトは 128 点とした。得られた対数振幅スペクトログラムは事前学習されたモデルの入力形状に合わせるため、低域から高域方向に 224 点、時間経過方向に 224 点の正方形で切り取った。

表2 FSDKaggle2018 を用いた分類正解率 (%)

	モデル	ResNet50		VGG16		
		Fine-Tuning の有無	なし	あり	なし	あり
直接	(a) 複製		85.6	<b>88.7</b>	80.1	<b>88.1</b>
	(b) 動的特徴量		<b>72.4</b>	46.4	<b>85.6</b>	79.5
	(c) カラー画像化		83.3	<b>90.3</b>	<b>89.6</b>	88.8
	(d) 帯域分割		86.5	<b>88.2</b>	82.2	<b>88.1</b>
間接	(e) 畳み込み (2 層)		82.8	<b>89.8</b>	<b>87.6</b>	81.0
	(f) 畳み込み (1 層)		86.0	<b>86.9</b>	86.9	<b>87.4</b>

なお手法 (d) のみ、情報量を確保するためにフレーム長を 1545 点、フレームシフトを 128 点として得た対数振幅スペクトログラムを同様に、低域から高域方向に 672 点、時間経過方向に 224 点で切り出し、周波数方向に 3 等分している。

本実験では ImageNet で事前学習された ResNet50 と VGG16 という 2 種類のネットワークに対してファインチューニングを行った。VGG16 は 13 層の畳み込み層と 2 層の全結合層及び出力層から構築されており、ResNet50 は層間の残差を学習することにより構造を深くすることを可能にしたネットワークで 49 層の畳み込み層と出力層で成り立っている。最適化アルゴリズムには Adam [25] を使用し、学習率を 0.0001、バッチサイズを 32、エポック数を 100 に設定し、全エポックの中から評価用データでの正解率が最良だったモデルを用いてテストを行った。すべての実験を 3 回行い、分類正解率の平均を実験結果として示した。分類正解率は少数点以下第 2 位を四捨五入している。比較手法は 3 章で述べた手法 (a) から (f) までの 6 種類である。手法 (c) ではスペクトログラムの値の低いものから高いものにかけて色相が青から赤に変化するように着色するため、Python 標準ライブラリである Matplotlib を用いてカラーマップを jet に設定することでカラー画像のスペクトログラムを生成した。

### 4.3 実験結果

表 1 に NSynth を使用した場合の実験結果を示す。実験結果より、ResNet50 と VGG16 を比較すると ResNet50 のほうが全体的に正解率が高いことがわかる。これは ResNet50 のほうが層が深いためモデルの表現能力が高いからだと考えられる。また、各モデルのファインチューニングの有無における性能を比較すると ResNet50 の動的特徴量以外はファインチューニングするほうが性能が高くなることが確認できる。手法ごとの違いを見るとカラー画像化が ResNet50 では一番性能が高く、VGG16 においては帯域分割が一番性能が高かった。カラー画像化は画像分類のタスクに入力情報を寄せていることから学習の親和性が高く、深い層の学習にも適していたと考えられる。一方、層の浅い VGG16 においては他のチャンネル変換法と比べて各チャンネルに含まれる調波構造が明確に表示されている帯域分割は特徴が捉えやすかったと考えられる。動的特徴量の性能が低いことに関しては、時系列情報の埋め込みがモデルの重みと合いにくく深い層では学習が十分に行えなかったことが原因だと考えられる。次に、間接的にチャンネル変換を行う手法について比較すると、層が浅いモデルではモデルの表現能力不足によりチャネ

ル変換が十分に学習出来なかったと考えられる。特に、3.2 節でも述べたとおり 1 層での畳込みは複製に近い情報となっており、精度も複製とほぼ等しいことから直接的なチャンネル変換とほぼ同じような特徴抽出しか出来ていないことが考えられる。

表 2 に FSDKaggle2018 を使用した場合の実験結果を示す。ResNet50 と VGG16 を比較すると動的特徴量以外は ResNet50 のほうが高い性能となっていることがわかる。また、各モデルのファインチューニングの有無における性能を比較すると ResNet50 では NSynth のときと同様に動的特徴量以外がファインチューニングするほうが性能が高くなるが、VGG16 では半分の手法でファインチューニングしないほうが高くなっている。FSDKaggle2018 は NSynth と比べてデータ量が少なく、かつ演奏音になっていることから複雑なモデル表現が必要であり、層の浅い VGG ではファインチューニングがうまく働かない場合があったと考えられる。また、手法 (e) 及び手法 (f) を比較すると、ResNet50 を用いると手法 (e) が、VGG16 を用いると手法 (f) が高い性能となっている。これは FSDKaggle2018 は NSynth に比べて難易度の高いデータセットであるため、層が深く表現力の高い ResNet50 ではうまく学習がすすみ、層が浅く表現力の低い VGG16 ではうまく学習できなかった可能性が考えられる。

## 5. ま と め

本研究では、楽器音分類のための画像分類ネットワークを用いたファインチューニングに関するチャンネル変換法について、チャンネル変換法の特徴を分析しつつ比較評価した。実験結果より、画像分類ネットワークでファインチューニングを行う際は元のタスクに近いカラー画像化によるチャンネル変換がもっとも性能が高くなった。今後の課題としては、他のチャンネル変換法の調査やカラー画像化の色の変化による調査、他のタスクでの評価などが挙げられる。

## 謝 辞

本研究の一部は JSPS 科研費 JP19K20271 及び JP20H00613 の助成を受けたものである。

## 文 献

- [1] 井本 桂右, "音響イベントと音響シーンの分析," 日本音響学会誌, Vol.74, No.4, p.198-207, 2018.
- [2] A. Mesaros, T. Heittola, T. Virtanen, "A multi-device dataset for urban acoustic scene classification." Proceedings of the detection and classification of acoustic scenes and events 2018 workshop (DCASE2018), 2018. November.
- [3] T. Heittola, A. Mesaros, T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," Proceedings of the detection and classification of acoustic scenes and events 2020 workshop (DCASE2020), 2020.
- [4] G. Dekkers, L. Vliegen, T.V. Waterschoot, B. Vanrumste, P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," Technical report, KU Leuven, 2018.
- [5] N. Turpault, R. Serizel, A.P. Shah, J. Salamon, "Sound event detection in domestic environments with weakly labeled data and sound-scene synthesis," Workshop on detection and classification of acoustic scenes and events, New York City, United States, October. 2019.
- [6] K. Patil, M. Elhilali, 2015, "Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases," EURASIP J.Audio Speech Music Process, 2015.
- [7] P.J. Donnelly, J.W. Sheppard, "Cross-Dataset validation of feature sets in musical instrument classification," In Proc. IEEE ICDMW, 94-101, 2015.
- [8] Tadanobu Inoue, P. Vinayavekhin, S. Wang, D. Wood, N. Greco, Ryuki Tachibana, "Domestic activities classification based on CNN using shuffling and mixing data augmentation," DCASE2018 Challenge, September. 2018.
- [9] H. Chen, Z. Liu, Z. Liu, P. Zhang, Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," June. 2019.
- [10] K. Koutini, H. Eghbal-zadeh, G. Widmer, "Receptive-field-regularized CNN variants for acoustic scene classification," June. 2019.
- [11] M. Dąbrowski, T. Michalik, O. Polska, "How effective is transfer learning method for image classification," FedCSIS, 2017.
- [12] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, "ImageNet: A large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition," CVPR, pp.248-255, Miami, 2009.
- [13] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, "Convolutional neural networks for medical image analysis: full training or fine tuning?," IEEE Transactions on Medical Imaging, vol.35, no.5, pp.1299-1312, May. 2016.
- [14] P. Marcelino, "Transfer learning from pre-trained models, Towards Data Science," <https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751>, May. 2020.
- [15] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," DCASE2018 Challenge, September. 2018.
- [16] S. Zhang, T. Huang, W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," IEEE Transactions on Multimedia, vol.20, no.6, pp.1576-1590, June. 2018.
- [17] M.N. Stolar, M. Lec, R.S. Bolia, M. Skinner, "Real time speech emotion recognition using RGB image classification and transfer learning," ICSPCS, 2017.
- [18] A. Guzhov, F. Raue, J. Hees, A. Dengel, "ES-ResNet: environmental sound classification based on visual domain models," arXiv preprint arXiv:2004.07301, 2020.
- [19] S. Adapa, "Urban sound tagging using convolutional neural networks," DCASE2019 Challenge, September. 2019.
- [20] H. Kaiming, Z. Xiangyu, R. Shaoqing, S. Jian, "Deep residual learning for image recognition," CVPR, June. 2016.
- [21] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," ICLR, 2015.
- [22] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," 2017.
- [23] E. Fonseca, M. Plakal, F. Font, D.P.W. Ellis, X. Favory, J. Pons, X. Serra, "General-purpose tagging of freesound audio with audioSet labels: task Description, dataset, and baseline," Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 2018.
- [24] U. Shukla, U. Tiwari, V. Chawla and S. Tiwari, "Instrument classification using image based transfer learning," 2020 5th International Conference on Computing, Communication and Security (ICCCS), pp. 1-5, 2020.
- [25] D.P. Kingma, J.L. Ba, "Adam: a method for stochastic optimization," ICLR 2015, 2015.